

1 METHODS AND APPARATUS FOR CONTROLLING THE FLOW OF MULTIPLE  
2 SIGNAL SOURCES OVER A SINGLE FULL DUPLEX ETHERNET LINK  
3

4 BACKGROUND OF THE INVENTION  
5

6 1. Field of the Invention

7 This invention relates broadly to telecommunications.  
8 More particularly, this invention relates to methods and  
9 apparatus for controlling the flow of multiple SONET signal  
10 streams over a single full duplex ETHERNET link.  
11

12 2. State of the Art

13 The TRANSWITCH ETHERMAP-12 is a highly integrated OC-  
14 12 mapper for carrying ETHERNET traffic over SONET/SDH  
15 networks utilizing Virtual Concatenation (VCAT). It  
16 supports STM-4/STS-12/STS-12c rates using a parallel  
17 telecom bus operating at 77.76 MHz. The device supports up  
18 to eight 10 Mbps or 100 Mbps ETHERNET ports using the SMII  
19 interface standard or a single Gigabit (1,000 Mbps)  
20 ETHERNET port using the GMII interface standard.  
21

22 When the ETHERMAP-12 is operated in the SMII mode,  
23 eight FIFOs are provided, one for each ETHERNET port, each

1 ETHERNET port being associated with one SONET port, virtual  
2 port or virtual concatenated group (VCG). Each FIFO has a  
3 high and a low threshold point which are associated with  
4 defined Xon (transmit data on) and Xoff (transmit data off)  
5 conditions. When a FIFO exceeds the Xoff threshold, a  
6 pause frame is generated. The pause duration is  
7 programmable and is identified in the pause frame. When  
8 the FIFO re-crosses the Xon threshold, a pause frame with a  
9 very short pause duration is generated. When operated in  
10 the SMII mode, the ETHERMAP-12 can support an OC-3 ring  
11 (155 Mbps) by combining two of the eight ETHERNET ports.

12

13 When the ETHERMAP-12 is operated in Gigabit mode, a  
14 single FIFO is provided for the single Gigabit ETHERNET  
15 port. In this mode, the ETHERMAP-12 supports a single OC-  
16 12 ring (622 Mbps). It would be desirable to multiplex a  
17 plurality of SONET ports, virtual ports or virtual  
18 concatenated groups (VCGs) over the single Gigabit ETHERNET  
19 link. For example, it would be desirable to support  
20 multiple OC-3 rings in the Gigabit mode of the ETHERMAP-12.

## SUMMARY OF THE INVENTION

It is therefore an object of the invention to provide methods and apparatus for multiplexing multiple signal sources over a single full duplex ETHERNET link.

It is another object of the invention to provide methods for multiplexing multiple signal sources over a single full duplex ETHERNET link using existing equipment.

It is a further object of the invention to provide methods for multiplexing multiple signal sources over a single full duplex ETHERNET link using an ETHERMAP-12 chipset.

It is also an object of the invention to provide methods for multiplexing a plurality of SONET ports over a single full duplex ETHERNET link using existing equipment.

It is an additional object of the invention to provide methods for multiplexing a plurality of SONET signal sources over a single full duplex gigabit ETHERNET link.

1       It is still another object of the invention to provide  
2 methods for multiplexing a plurality of SONET signal  
3 sources over a single full duplex gigabit ETHERNET link  
4 using existing equipment.

5  
6       It is yet another object of the invention to provide  
7 methods and apparatus which provide flow control for a  
8 multiplexed plurality of signal sources over a single  
9 ETHERNET link.

10  
11       It is still another object of the invention to provide  
12 methods and apparatus for controlling the flow of multiple  
13 signal sources over a single ETHERNET link.

14  
15       In accord with these objects, which will be discussed  
16 in detail below, methods for providing flow control  
17 according to the invention include receiving multiple data  
18 streams over a single ETHERNET link, associating a buffer  
19 with each data stream, putting received data into the  
20 appropriate buffer, monitoring the fullness of the buffers,  
21 and transmitting a PAUSE frame to the source of the data  
22 streams, the PAUSE frame indicating the fullness of each  
23 buffer. The methods for controlling the flow according to

1 the invention include reading the PAUSE frame and halting  
2 the transmission of data destined for a congested buffer(s)  
3 until a subsequent PAUSE frame is received which indicates  
4 that the congested buffer(s) has become decongested.  
5 Apparatus for performing the methods are also provided.

6  
7 Additional objects and advantages of the invention  
8 will become apparent to those skilled in the art upon  
9 reference to the detailed description taken in conjunction  
10 with the provided figures.

11

#### 12 BRIEF DESCRIPTION OF THE DRAWINGS

13

14 Fig. 1 is a high level schematic diagram illustrating  
15 bi-directional operation of the invention;

16

17 Fig. 2 is a high level schematic diagram illustrating  
18 the details of flow control in one direction;

19

20 Fig. 3A is an illustration of a prior art PDU MAC  
21 Encapsulation format;

1        Fig. 3B is an illustration of a modified PDU MAC  
2        Encapsulation format according to a first embodiment of the  
3        invention;

4  
5        Fig. 3C is an illustration of a modified PDU MAC  
6        Encapsulation format according to a second embodiment of  
7        the invention;

8  
9        Fig. 3D is an illustration of a modified PDU MAC  
10       Encapsulation format according to a third embodiment of the  
11       invention;

12  
13       Fig. 4 is a more detailed illustration of the modified  
14       PDU MAC Encapsulation format according to the first  
15       embodiment of the invention;

16  
17       Fig. 5 is a more detailed illustration of the modified  
18       PDU MAC Encapsulation format according to the second  
19       embodiment of the invention;

20  
21       Fig. 6 is a more detailed illustration of the modified  
22       PDU MAC Encapsulation format according to the third  
23       embodiment of the invention;

1        Fig. 7 is a schematic illustration of a modified pause  
2 frame according to the invention;

3

4        Fig. 8 is a schematic illustration of pause frame  
5 generation using timers according to the invention; and

6

7        Fig. 9 is a schematic illustration of an alternative  
8 modified pause frame according to the invention.

9

#### 10        DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

11

12        Turning now to Fig. 1, the invention is illustrated in  
13 a high level form with reference to a Layer 2/3 NPU  
14 (network processing unit) MAC (machine access control)  
15 client 10 and an ETHERNET over SONET (EoS) framer/mapper  
16 MAC client 12. The clients 10, 12 are coupled to each  
17 other by a full duplex gigabit ETHERNET link 14. As  
18 illustrated in Fig. 1, traffic from left to right is  
19 considered to be upstream to a plurality of SONET ports or  
20 VCGs 16. Thus, the client 10 is provided with a plurality  
21 of upstream transmit buffers 10a (one for each data  
22 stream), a plurality of downstream receive buffers 10b, an  
23 upstream transmit addressing and scheduling module 10c, a

1 downstream receive addressing module 10d, and a receive  
2 congestion monitor 10e. Similarly, the client 12 is  
3 provided with a plurality of downstream transmit buffers  
4 12a, a plurality of upstream receive buffers 12b, a  
5 downstream transmit addressing and scheduling module 12c,  
6 an upstream receive addressing module 12d, and a receive  
7 congestion monitor 12e.

8  
9       According to the invention, the upstream transmit  
10 addressing and scheduling module 10c receives a packet from  
11 one of the buffers 10a and encapsulates it in a modified  
12 MAC frame which includes an identification of which one of  
13 the destination ports 16 should receive the packet. The  
14 upstream receive addressing module 12d receives the MAC  
15 frame, decapsulates the packet and places the packet in one  
16 of the buffers 12b which corresponds to the destination  
17 port. The upstream receive congestion monitor 12e monitors  
18 the fullness of the buffers 12b and, when appropriate,  
19 generates a modified PAUSE control frame. The control  
20 frame is transmitted downstream by the downstream transmit  
21 addressing and scheduling module 12c, is received by the  
22 downstream receive addressing module 10d, and is used to  
23 control the upstream transmit addressing and scheduling



1 module 10c. In particular, the control frame causes the  
2 upstream transmit addressing and scheduling module 10c to  
3 cease transmitting packets destined for the congested  
4 buffer 12b. When congestion is relieved, a control frame  
5 indicating so is transmitted to the downstream receive  
6 addressing module 10d which causes the upstream transmit  
7 addressing and scheduling module 10c to resume transmitting  
8 packets to the decongested buffer.

9

10 Data traffic flow in the down stream direction  
11 operates in a similar manner. Packets received from the  
12 SONET ports 16 are placed in downstream transmit buffers  
13 12a (one for each SONET port). These packets are each  
14 encapsulated by the downstream transmit addressing and  
15 scheduling module 12c in a modified MAC frame which  
16 includes an identification of which one of the destination  
17 buffers 10b should receive the packet. The downstream  
18 receive addressing module 10d receives the MAC frame,  
19 decapsulates the packet and places the packet in one of the  
20 buffers 10b. The downstream receive congestion monitor 10e  
21 monitors the fullness of the buffers 10b and, when  
22 appropriate, generates a modified PAUSE control frame. The  
23 control frame is transmitted upstream by the upstream

1 transmit addressing and scheduling module 10c, is received  
2 by the upstream receive addressing module 12d, and is used  
3 to control the downstream transmit addressing and  
4 scheduling module 12c. In particular, the control frame  
5 causes the downstream transmit addressing and scheduling  
6 module 12c to cease transmitting packets destined for the  
7 congested buffer 10b. When congestion is relieved, a  
8 control frame indicating so is transmitted to the upstream  
9 receive addressing module 12d which causes the downstream  
10 transmit addressing and scheduling module 12c to resume  
11 transmitting packets to the decongested buffer.

12

13 Fig. 2 shows more detail of the upstream flow of  
14 ETHERNET traffic flowing from left to right with the PAUSE  
15 Frame flow control flowing from right to left. Independent  
16 flow control loops are shown in the downstream direction by  
17 dotted lines. Depending on the application, flow control  
18 in both directions may not be necessary. For example, flow  
19 control may not be required for traffic flowing from an EoS  
20 Framer/Mapper 12 to an NPU device 10. Traffic management  
21 with large buffering capacity is typically located in the  
22 NPU 10 and the EoS Framer 12 functions as a streaming  
23 device.

1       Upstream packets from buffers 10a-1...10a-n are  
2 multiplexed by the transmit addressing and scheduling block  
3 10c. The packets are encapsulated in MAC frames containing  
4 an address tag by the MAC block 10f. The MAC frames are  
5 decapsulated by the MAC block 12f. Details of the MAC  
6 blocks can be found in IEEE Standard 802.3-2002, Section  
7 One, paragraphs 2 through 4.4.3, pages 33-82, the complete  
8 disclosure of which is hereby incorporated by reference  
9 herein. The address tag is removed from the MAC frame by  
10 the receive addressing block 12d which recovers the  
11 original Ethernet packet and stores the PDUs in the  
12 appropriate Virtual Port FIFOs 12b-1...12b-n based on the  
13 address tag. Each Virtual Port FIFO is associated with a  
14 SONET Port or VCG 16-1...16-n. The receive congestion  
15 monitor 12e monitors the fill levels of the Virtual Port  
16 FIFOs 12b-1...12-b-n and sends PAUSE Control Frames when  
17 required to throttle the arrival of data frames for Virtual  
18 Port FIFOs that are nearing their buffering capacity. The  
19 PAUSE Control Frames are generated by the congestion  
20 monitor 12e and sent to the remote transmit addressing and  
21 scheduling block 10c in the remote ETHERNET client 10.  
22 Scheduling and transmission of packets to congested FIFOs  
23 is temporarily halted based on the content of the PAUSE

1 Control Frames. As the Virtual Port FIFO becomes  
2 decongested, another PAUSE Control Frame is sent to resume  
3 scheduling for the affected FIFO.

4  
5 An important feature of the invention is the modified  
6 MAC frame. Fig. 3A shows a standard gigabit ETHERNET MAC  
7 frame. It includes a six byte destination address DA, a  
8 six byte source address SA, a two byte Type/Length  
9 indicator, a variable length Payload, and a four byte frame  
10 check sum FCS.

11  
12 According to a first, though not presently preferred,  
13 embodiment of the invention, a two byte address and parity  
14 indicator is pre-pended to the MAC frame as shown in Fig.  
15 3B. According to this embodiment, the address tag is nine  
16 bits LSB justified and optionally protected by an odd  
17 parity bit. This format is not preferred because it is not  
18 "well formed" and will cause errors at the receiving MAC  
19 interface unless the interface is programmed to expect the  
20 extra two bytes at the start of each frame. Fig. 4 more  
21 clearly illustrates the arrangement of address bits for  
22 this embodiment. Bits 15-10 are set to ones so that the  
23 frame does not appear as a MAC control frame. Bit 9 is the

1 optional odd parity bit and bits 8-0 are the virtual port  
2 number.

3  
4 According to a second, and presently preferred,  
5 embodiment, the address tag is mapped onto a standard (IEEE  
6 802.1Q) VLAN stacked label. In this embodiment, which is  
7 illustrated in Fig. 3C, the frame check sum bytes reflect  
8 the additional VLAN fields. Fig. 5 more clearly  
9 illustrates the arrangement of bits for this embodiment.  
10 Bits 15-9 are set to zeros and bits 8-0 are used for the  
11 virtual port number.

12  
13 A variant of the embodiment is shown in Fig. 3D  
14 wherein an existing VLAN ID in the frame is mapped to a  
15 virtual port address. This variant can only be used where  
16 there is a 1:1 correspondence between VLAN IDs and virtual  
17 port addresses. Both of these addressing methods (i.e. the  
18 embodiments of Figs. 3C and 3D) are well-formed and will  
19 not produce errors at the receiving MAC interface. Fig. 6  
20 more clearly illustrates the bits of the VLAN ID which are  
21 mapped to a virtual port address. Bits 15-12 are ignored  
22 and bits 11-0 are mapped to a virtual port number.

1        Another important feature of the invention is the use  
2   of a modified PAUSE control frame. Fig. 7 illustrates a  
3   modified PAUSE control frame according to one embodiment of  
4   the invention. Like all MAC frames, the PAUSE control  
5   frame has a header which includes a six byte destination  
6   address field, a six byte source address field, and a two  
7   byte length/type field. A two byte MAC control opcode  
8   follows the header and the control frame message follows.  
9   Since the minimum frame size for gigabit ETHERNET is 64  
10   octets, a minimum of 44 octets is available for the control  
11   frame message which is followed by a four byte frame check  
12   sum. According to the presently preferred embodiment of  
13   the invention, the minimum frame size is used for the PAUSE  
14   control frame. Following the MAC control opcode, a sixteen  
15   bit PAUSE timer field contains one of two timer values and  
16   following the timer field are three hundred thirty-six bits  
17   corresponding to three hundred thirty-six virtual ports.  
18   Each of these bits indicates the state of the associated  
19   port, e.g. 0=XON, 1=XOFF.

20

21        According to the presently preferred embodiment of the  
22   invention, four programmable timer values are provided.  
23   The first is the "Pause\_Time\_Value", a 16-bit Read/Write

1 timer value that is configurable from the host interface  
2 and is one of the two values assumed by the PAUSE timer  
3 field in the PAUSE control frame. The other value assumed  
4 by the PAUSE timer field is zero.

5  
6 The second timer value is the "Pause\_Delay\_Timer", a  
7 16-bit Read/Write timer that is configurable from the host  
8 interface. Each timer tick is in units of 512 bit times on  
9 the gigabit ETHERNET interface. The Pause\_Delay\_Timer  
10 represents an XON/XOFF transition "window" in which  
11 multiple virtual ports will have their state changes  
12 accumulated and sent in a single PAUSE Control Frame to the  
13 remote MAC client. A value of 1 indicates that a new PAUSE  
14 Control Frame is allowed to be generated every 512 bit  
15 times or 512ns. A value of 65535 indicates that a new  
16 PAUSE Control Frame is allowed to be generated every  
17 33.5ms. Larger values limit the percentage of bandwidth  
18 that can be occupied by PAUSE Control Frames at the cost of  
19 increased latency. Use of this timer is OPTIONAL if the  
20 third timer, the "Pause\_Refresh\_Timer" is used to send  
21 periodic updates continuously.

22

1       The "Pause\_Refresh\_Timer" is a 16-bit Read/Write timer  
2   that is configurable from the host interface. Each timer  
3   tick is in units of 512 bit times on the gigabit ETHERNET  
4   interface. The Pause\_Refresh\_Timer represents a periodic  
5   refresh rate to the remote MAC client when there have been  
6   no transitions in virtual port XON/XOFF states for the  
7   refresh time period. This timer is properly set to a value  
8   that is slightly lower than the Pause\_Time\_Value in order  
9   to guarantee that extended XOFF states are refreshed before  
10  the timer expires on the remote MAC client.

11

12       The fourth timer is the "Pause\_Delay\_Timer\_Tx". This  
13  is a 16-bit timer that is updated with the timer value  
14  supplied in the received PAUSE Control Frames from the  
15  remote MAC client. When at least one port is in the XOFF  
16  state, a value of one is expected to be present in at least  
17  one of the PAUSE state bit fields. When all ports are in  
18  the XON state, a value of zero is expected to be present in  
19  all of the PAUSE state bit fields. The timer begins  
20  decrementing at the rate of one tick every 512 bit times  
21  after it is updated from the PAUSE Control Frame. When the  
22  timer reaches zero, all virtual ports return to the XON  
23  state.



1        Fig. 8 illustrates an example of the flow of PAUSE  
2        control frames from the framer/mapper to the NPU. Reading  
3        Fig. 8 from top to bottom and right to left, the sequence  
4        begins with the initial condition where all ports are in  
5        the XON state. At time  $t_1$ , a PAUSE control frame  
6        indicating all ports XON is sent (by the congestion monitor  
7        12e shown in Figs. 1 and 2) with the Pause Timer = 0000 and  
8        all of the Pause State bits = 0. After the expiration of  
9        the Pause refresh time  $t_2 - t_1$ , the same control frame is  
10       sent (by the congestion monitor 12e shown in Figs. 1 and 2)  
11       at time  $t_2$ . At some time following  $t_2$ , virtual ports 1 and  
12       3 become congested. Before sending a control frame  
13       indicating congestion, the congestion monitor 12e (shown in  
14       Figs. 1 and 2) waits for the expiration of the Pause Delay  
15       Timer which is reset each time the refresh frame is sent.  
16       The Pause Delay Timer provides a window within which  
17       congestion may be cleared without putting a port in an XOFF  
18       state. Upon the expiration of the Pause Delay Timer at  $t_3$ ,  
19       a PAUSE control frame is sent (by the congestion monitor  
20       12e shown in Figs. 1 and 2) indicating that ports 1 and 3  
21       should be put in XOFF state with the Pause Timer = Pause  
22       Time Value. Transmission of packets destined for ports 1  
23       and 2 is temporarily halted.

1        So long as there is no change in the congestion status  
2 of the virtual ports, the Pause Refresh Timer is allowed to  
3 expire before another control frame is sent. Thus, at t4,  
4 the same control frame that was sent at t3 is sent again,  
5 indicating that conditions are the same as at t3. At some  
6 time following t4, port 1 becomes decongested and port 4  
7 becomes congested. Upon the expiration of the Pause Delay  
8 Timer, a new PAUSE control frame is sent at t5 indicating  
9 the new status of the ports and setting the Pause Timer to  
10 the Pause Time Value. Transmission of packets destined for  
11 port 1 is resumed and transmission of packets destined for  
12 port 4 is temporarily halted.

13

14        At some time following t5, port 3 becomes decongested.  
15 Thus, upon the expiration of the Pause Delay Timer, a new  
16 PAUSE control frame is sent at t6 indicating the new status  
17 of the ports and setting the Pause Timer to the Pause Time  
18 Value. Transmission of packets destined for port 3 is  
19 temporarily halted. Following time t6 until the expiration  
20 of the Pause Refresh Timer at t7, there is no change in the  
21 congestion status of the ports. Therefore, at t7, the same  
22 control frame that was sent at t6 is sent again.

1       At some point following t7, port 4 becomes  
2   decongested. Thus, upon the expiration of the Pause Delay  
3   Timer, a new PAUSE control frame is sent at t8 indicating  
4   the new status of the ports (all decongested) and setting  
5   the Pause Timer to 0000. Transmission is resumed for all  
6   ports. Following time t8 until the expiration of the Pause  
7   Refresh Timer at t9, there is no change in the congestion  
8   status of the ports. Therefore, at t9, the same control  
9   frame that was sent at t8 is sent again.

10

11       In applications where End-to-End PAUSE Control is not  
12   allowed to be transported across the Sonet/SDH network,  
13   PAUSE frames arriving from the Sonet/SDH network shall be  
14   discarded and no further action taken. The PAUSE frames  
15   sent over the ETHERNET interface shall reflect only the  
16   local Upstream Rx buffer congestion.

17

18       In applications where End-to-End PAUSE Control is  
19   allowed, the Rx Congestion Monitor (12e shown in Figs. 1  
20   and 2) must perform a PAUSE reconciliation between the  
21   remote pause condition arriving from the Sonet/SDH network  
22   and the local upstream Rx buffer congestion managed  
23   locally. When the Sonet/SDH side is not asserting a pause

1 condition, PAUSE frames sent on the ETHERNET interface  
2 follow the mux-mode format and scheme described above with  
3 reference to Figs. 7 and 8. When the Sonet/SDH side is  
4 asserting a pause, the "Rx Congestion Monitor" function  
5 must reconcile both local and remote congestion conditions  
6 on a per virtual port basis (VCG). To perform  
7 reconciliation, a Remote\_Pause\_Timer (per VCG) must be  
8 maintained reflecting the XOFF period identified in the  
9 Pause\_Time field requested from the remote side across the  
10 Sonet/SDH network. This timer is decremented using a time  
11 quantum of 1 tick per 512 bit times for a 10Mbit interface  
12 (i.e. 51.2  $\mu$ s per tick). As long as the remote side is  
13 asserting/refreshing XOFF, the mux-mode PAUSE frames sent  
14 on the ETHERNET interface will be in the XOFF state. When  
15 the remote side is asserting the XON state, i.e.  
16 Remote\_Pause\_Timer (per VCG) decrements or is set to zero,  
17 PAUSE frames sent on the ETHERNET interface should reflect  
18 the state of local congestion maintained for the Upstream  
19 Rx buffer for the virtual port/VCG.

20

21 Fig. 9 illustrates an alternative Pause Control Frame  
22 where each port is allocated two bits in the message  
23 payload to identify XON, XOFF, and NO-CHANGE state. The

1 NO-CHANGE state allows better control of the pause delays  
2 imposed on virtual ports at the time the delay is imposed.  
3 For example, when virtual ports 1, 3, and 4 require  
4 backpressure during interval "X", a single Pause frame is  
5 generated which tells each of these ports to delay by the  
6 time specified in the Pause Timer. If during interval "Y",  
7 some time later, virtual ports 6 and 7 also need to be  
8 backpressured, VP's 1,3, and 4 are already into their  
9 countdown period and are assigned NO-CHANGE along with all  
10 other ports that were not in a backpressure state. VP's 6  
11 and 7 begin a new backpressure time period with an initial  
12 (high) Pause\_Timer.

13

14 This implementation of the PAUSE control frame is not  
15 preferred for two reasons. First, it is advantageous to  
16 use a single bit encoding for XON/XOFF backpressure to have  
17 a simple and compact representation within the Pause Frame  
18 payload. Second, since XOFF is nominally asserted for the  
19 maximum Pause\_Timer, this can be continued for all ports in  
20 the congestion state until Rx FIFO congestion is alleviated  
21 and XON is asserted on all ports below the XON threshold.

22

1       There have been described and illustrated herein  
2 several embodiments of methods and apparatus for  
3 controlling the flow of multiple signal sources over a  
4 single full duplex ETHERNET link. While particular  
5 embodiments of the invention have been described, it is not  
6 intended that the invention be limited thereto, as it is  
7 intended that the invention be as broad in scope as the art  
8 will allow and that the specification be read likewise.  
9 Thus, while the invention has been described with reference  
10 to gigabit ETHERNET, it will be appreciated that the  
11 invention could be applied to ETHERNET links of different  
12 bandwidth as well. In addition, while particular types of  
13 modified MAC frames have been disclosed, it will be  
14 understood other types of modified MAC frames might be able  
15 to obtain similar results. Also, while a particular  
16 modified PAUSE control frame is preferred, it will be  
17 recognized that other formats may be able to obtain similar  
18 results if designed with the present disclosure in mind.  
19 It will therefore be appreciated by those skilled in the  
20 art that yet other modifications could be made to the  
21 provided invention without deviating from its spirit and  
22 scope as claimed.